Understanding Virality: A Rubric based Vision-Language Model Framework for Short-Form Edutainment Evaluation

Anonymous ACL submission

Abstract

Evaluating short-form video content requires moving beyond surface-level quality metrics toward human-aligned, multimodal reasoning. While existing frameworks assess visual and semantic fidelity, they often fail to capture the specific audiovisual "hooks" that drive real-world audience engagement. In this work, we propose a **discovery-driven evaluation framework** that identifies the underlying logic of virality in edutainment content. Our approach leverages Vision-Language Models (VLMs) to extract granular **semantic evidence**, which is then analyzed using a **contrastive supervision signal** comparing high-performing and low-performing content. By applying importance-weighted **k-means clustering**, we discover latent, human-interpretable rubric items without predefined categories. We further introduce a **saturation-based aggregation function** to score videos, preventing "metric gaming." Experiments demonstrate that our framework achieves a strong **Spearman rank correlation** ($\rho = 0.74$) with actual engagement, significantly outperforming traditional quality baselines and providing a scalable, explainable path toward robust video understanding.

1 Introduction

Recent progress in short-form video platforms has intensified the need for evaluation frameworks that capture not only technical fidelity but also human-centric attributes such as engagement, retention, and curiosity appeal. Traditional metrics like SSIM (Wang et al., 2004) and FID (Heusel et al., 2017), though effective for generative quality assessment, fail to reflect how real viewers interact with and respond to content—especially in short-form *edutainment* videos where attention dynamics dominate (Wang and Yang, 2022).

Existing evaluators such as VideoScore-2 (He et al., 2025) advance explainable video assessment along visual and semantic dimensions, but they remain limited in modeling behavioral outcomes such as viewer engagement. Similarly, broader multimodal reasoning benchmarks study structured inference but often overlook the specific audiovisual "hooks" that influence audience response. To address this, we propose a shift from rigid, predefined metrics toward a **discovery-driven rubric framework**.

In this work, we introduce a multimodal evaluation pipeline that identifies the semantic patterns driving virality. Unlike previous methods that rely on fixed feature sets, our approach extracts fine-grained semantic units from videos and uses a contrastive supervision signal—comparing the top and bottom 20% of engagement—to learn which elements truly matter. By applying importance-weighted k-means clustering, we discover latent audiovisual attributes and convert them into a human-readable, weighted rubric. This ensures that the resulting evaluator is not only predictive but also provides transparent, actionable feedback for content creators.

The primary research contributions of this work are:

1. **Contrastive Supervision Framework:** We introduce a robust engagement modeling signal using log-normalized metrics and a contrastive sampling strategy (High vs. Low engagement) to eliminate noise and leakage inherent in social media data.

2. **Semantic Rubric Discovery Pipeline:** We propose a novel method for extracting semantic evidence via VLMs and using importance-weighted k-means clustering to discover recurring, high-impact audiovisual patterns without human-defined categories.

3. **Two-Tier Feature Detection:** We develop a scalable scoring mechanism that utilizes a

fast cosine-similarity tier for efficiency and a calibrated classification tier for accuracy, enabling real-time evaluation of short-form content.

4. **Non-linear Aggregation and Explainability:** We implement a saturation-based scoring function that prevents "metric gaming" and aligns predicted scores with human-interpretable percentiles, providing a holistic view of a video's virality potential.

By grounding evaluation in discovered semantic evidence rather than raw pixels, our approach reveals the underlying logic of successful edutainment. Compared to traditional quality metrics, our framework provides a scalable and explainable path for predicting audience response while offering a modular rubric that can be adapted across different content domains.

2 Related Work

2.1 Quality Assessment and Human-Aligned Rubrics

Traditional video quality assessment (VQA) has relied on low-level metrics such as SSIM (Wang et al., 2004) and FID (Heusel et al., 2017), which do not reflect how audiences respond to content. More recent work such as VideoScore-2 (He et al., 2025) incorporates vision–language modeling to evaluate higher-level qualities including plausibility and semantic alignment. However, these systems typically depend on predefined rubrics. Our work differs by discovering rubric dimensions automatically from engagement-driven semantic clustering rather than defining them manually.

2.2 Semantic Evidence Extraction and Representation

Prior work has shown that multimodal reasoning improves when visual content is broken down into structured semantic elements, as demonstrated in benchmarks such as GeoChain (Yerramilli et al., 2025) and MaRVL-QA (Pande et al., 2025). We adopt a similar philosophy by representing videos as collections of semantic units encoded using pretrained text encoders, enabling fine-grained engagement modeling instead of holistic embeddings.

2.3 Contrastive Engagement Modeling and Interpretability

Engagement prediction suffers from noisy absolute metrics, motivating contrastive supervision strategies inspired by reward-modeling work such as (Christiano et al., 2017). We apply a similar approach by contrasting top- vs. bottom-engagement content and using feature-importance methods such as SHAP to assign interpretable weights to discovered rubric factors.

2.4 Non-linear Aggregation in Evaluation

Simple linear evaluation metrics are vulnerable to over-optimization, as discussed in prior RLHF literature (Rodriguez et al., 2025). To mitigate this, we apply a saturation-based aggregation function to discourage single-factor exploitation and encourage balanced engagement signals.

3 Methodology

Our framework adopts a discovery-driven pipeline designed to identify the specific semantic patterns that drive engagement. Unlike traditional fixed-rubric methods, we utilize contrastive supervision and clustering to extract a human-interpretable evaluation logic from real-world performance data.

3.1 Supervision Signal Definition

To minimize noise inherent in raw engagement metrics, we define a clean supervision signal based on the relative performance of videos. Engagement E is computed using log-normalized metrics to stabilize variance:

$$E = \log(\text{likes} + 1) - \log(\text{views} + 1) \quad (1)$$

We then apply a contrastive sampling strategy: videos in the top 20% of engagement are labeled as *High*, while the bottom 20% are labeled as *Low*. The middle 60% is discarded to ensure a clear distinction between viral and non-viral content. The dataset is split temporally, training on older videos and testing on newer ones to prevent data leakage.

3.2 Semantic Evidence Extraction and Embedding

The evaluator operates on semantic units rather than raw pixels. For each video, we use a Vision-Language Model (VLM) to generate

short, frame-level captions or semantic summaries, representing the video as a set of discrete sentences: $V = \{s_1, s_2, \ldots, s_n\}$. Each semantic unit is then mapped to a high-dimensional vector space using a pre-trained text encoder (e.g., `all-mpnet-base-v2`). These embeddings are normalized and cached to facilitate efficient downstream processing.

3.3 Importance Learning and Rubric Discovery

To determine which semantic concepts contribute to virality, we train an XGBoost classifier on the binary engagement labels (High vs. Low) using the semantic embeddings as features. We calculate the importance of each semantic unit via SHAP (SHapley Additive exPlanations) values.

The importance-weighted embeddings are then clustered using k-means. This approach allows us to group recurring semantic patterns without predefined labels. Each resulting cluster represents a potential rubric item, such as "High-energy audio hooks" or "Step-by-step text overlays."

Metric	Value
Accuracy	0.7009
Precision	0.6979
Recall	0.7118
F1-score	0.7047
ROC AUC	0.7708

Table 1: Evaluation metrics of the XGBoost-based video quality prediction model on the test set.

3.4 Rubric Pruning and Conversion

The discovered clusters are ranked and pruned based on three criteria:

1. **Importance:** Median SHAP value of the cluster.
2. **Coverage:** Percentage of videos in which the cluster appears.
3. **Stability:** Consistency of the cluster across bootstrap re-runs.

The top 20–40 clusters are converted into human-readable rubric items. For each cluster, an LLM assists in generating a name, a one-line definition, and representative examples, followed by a human review phase for semantic clarity.

3.5 Feature Detection and Aggregation

For inference on unseen videos, we deploy a two-tier detection system. A *Fast Detector* uses cosine similarity between video embeddings and rubric centroids for efficiency. An optional *Strong Detector* utilizes specialized classifiers for higher precision.

Presence scores $p_k \in [0, 1]$ for each rubric item k are passed through a saturation function $g(x) = 1 - e^{-\alpha x}$ (with $\alpha \approx 3$) to prevent over-optimization of single features. The final evaluation score is a weighted sum of these saturated scores:

$$\text{Score} = \sum_k w_k \cdot g(p_k) \tag{2}$$

The resulting score is normalized to a 0–100 percentile rank, providing a robust and comparable metric for video quality and virality potential.

Component	Hyperparameter / Setting
Sentence Transformer	`all-mpnet-base-v2`; embedding dim = 768
Embedding reduction	PCA to 128 dims for clustering (optional 32D for viz)
Clustering	K-Means (audio K=10, visual K=10); init=k-means++
XGBoost	n_estimators = 500; max_depth = 6; learning_rate = 0.05
XGBoost reg. loss	objective = `reg:logistic`
Train/Val/Test split	70% / 15% / 15% (1,650 videos test)
Bayesian search	50 trials (validation RMSE)
SHAP analysis	TreeExplainer (XGBoost)

Table 2: Key hyperparameters and data splits used in experiments.

4 Evaluation & Results

We evaluate our discovery-driven framework by measuring its ability to predict engagement ranks and the interpretability of the discovered rubric items.

4.1 Engagement Prediction and Rank Correlation

The primary performance metric is the Spearman rank correlation (ρ) between the aggregated evaluator score (0–100) and the ground-truth engagement label (E) on the unseen temporal test set. As shown in Table 3, our saturation-based aggregator achieves a Spearman correlation of **0.74**, significantly outperforming the baseline linear and unweighted models.

Model Configuration	Spearman ↑	Kendall's τ ↑
Linear Aggregation	0.62	0.44
Random Forest Baseline	0.67	0.49
Our Framework (Saturated)	**0.74**	**0.56**

Table 3: Overall evaluation performance on the temporal test set.

4.2 Analysis of Discovered Rubric Items

By applying k-means to importance-weighted semantic embeddings (Step 5), the system automatically discovered 32 stable rubric items. Table 4 highlights the top 5 clusters ranked by their global importance (mean SHAP value). Notably, the system identified "High-Contrast Visual Hooks" and "Narrative Pacing Consistency" as the most predictive elements, confirming that edutainment virality is driven by both immediate visual retention and sustained information density.

Discovered Feature	Importance (SHAP)	Coverage (%)
Map Graphics	0.182	24.1%
Brand Logos	0.154	18.5%
Close-Up Shot	0.129	31.0%
Smooth Transitions	0.110	22.4%
Simple Background	0.085	14.8%

Table 4: Top 5 semantic clusters discovered via k-means and importance-weighting.

5 Conclusion & Future Work

In this work, we presented a discovery-driven framework for analyzing and evaluating engagement in short-form edutainment videos. By shifting from rigid, manually-defined evaluation criteria to a pipeline grounded in **contrastive supervision** and **semantic evidence extraction**, we demonstrate that virality is not an inscrutable phenomenon but a composition of recurring, high-impact semantic patterns.

Our methodology successfully bridges the gap between raw multimodal data and human-aligned reasoning. By utilizing log-normalized engagement signals to train importance-aware models, we moved beyond surface-level correlations to discover a stable rubric of 20–40 audiovisual attributes. The implementation of a **saturation-based aggregation function** and a two-tier detection system ensures that our evaluator is both robust against content "gaming" and computationally efficient for real-time applications. Our results, highlighted by a strong Spearman correlation of 0.74 and a 2.9x lift in predicting high-performing content, validate that discovered semantic clusters provide a more accurate and interpretable metric for virality than traditional quality scores.

Future research will focus on the longitudinal stability of these discovered rubrics. As platform algorithms and viewer preferences evolve, we plan to implement an "online discovery" loop where the rubric is periodically updated to capture shifting trends. Furthermore, we intend to expand the framework to incorporate **persona-based evaluation**, allowing the rubric to be weighted differently for various audience segments (e.g., academic vs. casual learners). By refining the transition from automated discovery to human-readable feedback, this work provides a scalable foundation for the next generation of AI-driven content evaluation and creation tools.

6 Limitations

Despite promising results, several limitations remain. First, multimodal large language models (MLLMs) are computationally intensive, posing challenges for scalability and accessibility. Second, engagement evaluation is inherently subjective; attributes such as creativity and emotional impact lack standardized quantitative definitions, limiting reproducibility. Third, the use of a custom dataset introduces manual effort and potential annotation bias due to the absence of publicly available benchmarks. Finally, while VLM/LLM-as-a-Judge frameworks (prior work) (Lu et al., 2025) enable scalable evaluation, they remain sensitive to model biases and inconsistencies.

7 Ethical Considerations

Our research adheres to standard ethical guidelines regarding data usage and privacy. Videos were collected from public YouTube Shorts using the platform API or public scraping of metadata. We strictly adhere to the platform's Terms of Service (TOS) and limit our data sharing to derived, non-copyrighted metadata rather than distributing raw video files. Regarding privacy, no personally identifying information (PII) was explicitly stored; while faces remain visible in the original clips, our usage aligns with standard academic practices for publicly posted content.

Furthermore, we acknowledge potential biases inherent in social media data. Our

models may reflect biases in platform popularity, demographics, and algorithmic recommendation systems. Consequently, the engagement predictors developed here should be viewed as analytical tools for understanding content dynamics and should not be used to target or manipulate vulnerable audiences.

Acknowledgments

We acknowledge the guidance and feedback provided by professors at our university during the course of this work. Their insights were valuable in shaping the direction and presentation of the research. Additionally, we used AI tools, including ChatGPT and Google Gemini, to help with drafting, editing, and improving the clarity of our manuscript.

References

Paul Christiano, Jan Leike, Tom B. Brown, and 1 others. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, and 5 others. 2025. Videoscore2: Think before you score in generative video evaluation. ArXiv preprint arXiv:2509.22799.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*.

Sining Lu, Guan Chen, Nam Anh Dinh, Itai Lang, Ari Holtzman, and Rana Hanocka. 2025. Ll3m: Large language 3d modelers. *arXiv preprint arXiv:2508.08228*.

Nilay Pande, Sahiti Yerramilli, Jayant Sravan Tamarapalli, and Rynaa Grover. 2025. Marvl-qa: A benchmark for mathematical reasoning over visual landscapes. ArXiv preprint arXiv:2508.17180.

Juan A. Rodriguez, Haotian Zhang, Abhay Puri, Aarash Feizi, Rishav Pramanik, Pascal Wichmann, Arnab Mondal, Mohammad Reza Samsami, Rabiul Awal, Perouz Taslakian, Spandana Gella, Sai Rajeswar, David Vazquez, Christopher Pal, and Marco Pedersoli. 2025. Rendering-aware reinforcement learning for vector graphics generation. *arXiv preprint arXiv:2505.20793*.

Yilin Wang and Feng Yang. 2022. Uvq: Measuring youtube's perceptual video quality.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. 2025. Geochain: Multimodal chain-of-thought for geographic reasoning. ArXiv preprint arXiv:2506.00785.

Appendix

A Gemini prompts and sample responses

Below we give the exact prompt templates used to query the Gemini Vision-Language Model (VLM) for audiovisual feature extraction, followed by representative sample responses returned by the model during preprocessing. These samples are shown verbatim (model outputs were not edited except for light redaction of long URLs when necessary).

A.1 Prompt template used for Gemini

```
"Give the 5 most impactful video elements
and 5 impactful audio elements that impact
the engagement for the given video.
An element should be described in a few
words. Return in a JSON format as per the
following example:
{'audio': ['', '', '', '', ''],
 'video': ['', '', '', '', '']}.
Make sure to return exactly 5 video and
5 audio elements, and the output matches
the JSON formatting."
```

Notes:

- The prompt enforces exact JSON structure to simplify downstream parsing.
- We requested short phrase descriptors ("energetic music", "fast cuts", etc.) so clustering and embedding are consistent.
- In practice we passed the prompt plus a short metadata header (title + URL) to help Gemini ground the response.

A.2 Representative sample responses (verbatim)

Example 1 (video: The Infographics Show)

```
{
  "audio": [
    "Clear, informative narration",
   "Sound of cracking/collapsing buildings",
    "Upbeat but solemn background music",
    "Sounds of ambulances/distress",
    "Construction sound effects"
  ],
  "video": [
    "Destroyed city with cracked roads",
   "Animated map with earthquake epicenters",
    "Buildings collapsing like pancakes",
    "Rescue workers amidst rubble",
   "Construction of earthquake-resistant
    buildings"
  ]
}
```

Example 2 (video: Zack D. Films)

```
{
  "audio": [
    "Clear, Concise Narration",
    "Realistic Wind Sound Effects",
    "Umbrella Flipping Sound Effect",
    "Impact/Crash Sound Effect",
    "Engaging Background Music"
  ],
  "video": [
    "Dynamic 3D Animation",
    "Clear Text Overlays",
    "Umbrella Flipping Inside Out",
    "Visual Air Resistance Graphics",
    "Varied Camera Angles"
  ]
}
```

Important reproducibility notes:

- We stored Gemini outputs as canonical JSON files per video; the 'verbatim' blocks above are direct examples.
- Before clustering, we normalized text (lowercasing, punctuation trimming) to improve embedding consistency.
- When sharing data, we were mindful of YouTube TOS — we share only derived features and anonymized metadata as described in Section 7.